EXPRESS MAIL LABEL NO.:

(EV 304 737 783 US)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

METHOD AND SYSTEM FOR FAST LINK FAILOVER

INVENTOR(S):

ALBERT H. MITCHELL, JR.
PRITESH N. PATEL
APRIL CHOU
MAURICIO ARREGOCES
CHRISTOPHER SPAIN

Attorney Docket No.: **CIS0215US**

PREPARED BY:

CAMPBELL, STEPHENSON, ASCOLESE, LLP

4807 SPICEWOOD SPRINGS ROAD
BUILDING 4, SUITE 201
AUSTIN, TEXAS 78759

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] The present invention relates generally to communications networks and more particularly to a method and system for fast link failover.

DESCRIPTION OF THE RELATED ART

[0002] Companies today depend increasingly on the ability to quickly and reliably access data via communications networks. As the accessibility, reliability, and availability of such communications networks has become more important, a number of techniques to increase these factors have been developed. Redundancy is one such technique frequently used to minimize network downtime and increase the speed at which data may be accessed via a communications network. For example, redundant network links or connections are frequently used to couple a single network element or node, (e.g., a client, server, or other host data processing system, a communications network appliance, or a switch, router, hub, gateway or other redistribution point) to one or more communications networks or portions thereof (e.g., network segments). Such use of redundant network links with respect to network elements residing at the edge or terminating point of a communications network is known as "multi-homing" and such redundantly-linked network elements are said to be "multi-homed".

[0003] Fig. 1 illustrates a data processing system including multi-homed network elements. Data processing system 100 of the illustrated embodiment includes an upstream communications network portion 102, a primary switch 104a, a secondary switch 104b, and a number of multi-homed network elements (e.g., multi-homed endstations 106a, 106b...106n). Upstream communications network portion 102, including any of a number of network elements is coupled to primary switch 104a using a primary link 108a and to secondary switch 104b using a secondary link 108b. Multi-homed endstations 106a, 106b...106n are each similarly coupled (e.g., via a primary network interface card or host bus adapter) to primary switch 104a via one of a plurality of primary links 110a-110n and (e.g., via a secondary network interface card or host bus adapter) to secondary switch 104b via one of a plurality of secondary links 112a-112n as shown.

[0004] In operation, data is transmitted between multi-homed endstations 106 and upstream communications network portion 102 using primary links 110a-110n, primary

switch 104a, and primary link 108a. Following a failure of any of primary links 110a-110n, (e.g., due to failure of the physical link hardware, a primary network interface card, or primary switch 104a) one or more associated multi-homed endstations may failover to a corresponding secondary link 112a-112n by activating a network interface associated with the secondary link and deactivating a network interface associated with the failed primary link. Data is then transmitted between the multi-homed endstation which has failed over and upstream communications network portion 102 via an associated secondary link 112, a secondary switch 104b, and a secondary link 108b.

[0005] In a conventional communications network however, a failure of a link not directly attached to a network element (e.g., a failure of primary link 108a considered from the perspective of one or more of endstations 106a-106n) cannot be quickly detected. Traditionally, high-level system components (e.g., protocols, applications) have been utilized to detect the occurrence of such failures. For example, a high-level system component resident on an endstation 106 may use a timer to track a time period between data transfers associated with an upstream communications network portion or may use periodic link or connection status messages or "data units" to determine whether or not an upstream link failure has occurred.

[0006] The described techniques suffer from a number of significant shortcomings however. To account for ordinary communications network congestion and to avoid falsely declaring a link failure, the threshold time periods (and resultant latency) associated with the described techniques are typically multiple seconds or more. Additionally, such high-level system components frequently operate at an individual application or network element level. Consequently, failure of a link coupling an aggregating network element to an upstream communications network portion may not be simultaneously detected by all downstream network elements to which the aggregating network element is coupled.

SUMMARY OF THE INVENTION

[0007] Disclosed is a method and system for fast link failover. Using one or more embodiments of the present invention, network connectivity (e.g., data link layer connectivity) information is propagated, thereby enabling downstream network elements not immediately adjacent to the site of a link failure or directly coupled to a network element experiencing link failure to failover to alternate, redundant links such that the state of one or more connections or communications channels with the upstream portion(s) of a communications network may be preserved and the connection(s)/channel(s) may be maintained. According to one embodiment, a method is provided in which a failure of a first link between a network element and an upstream portion of a communications network is detected, and a second link between the network element and a downstream portion of the communications network is responsively disabled to maintain a communications channel between the downstream and upstream portions of the communications network using one or more alternate links.

[0008] The foregoing is a summary and thus contains, by necessity, simplifications, generalizations and omissions of detail; consequently, those skilled in the art will appreciate that the summary is illustrative only and is not intended to be in any way limiting. Other aspects, inventive features, and advantages of the present invention, as defined solely by the claims, will become apparent in the non-limiting detailed description set forth below.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] The present invention may be better understood, and its numerous features and advantages made apparent to those skilled in the art by referencing the accompanying drawings in which:

[0010] Fig. 1 illustrates a data processing system useable with one or more embodiments of the present invention;

[0011] Fig. 2 illustrates a data processing system including a primary switch network element according to an embodiment of the present invention;

[0012] Fig. 3 illustrates a link failure propagation process according to an embodiment of the present invention; and

[0013] Fig. 4 illustrates a data processing system including a primary Ethernet switch network element according to an embodiment of the present invention.

[0014] The use of the same reference symbols in different drawings indicates similar or identical items.

DETAILED DESCRIPTION

[0015] Although the present invention has been described in connection with one or more specific embodiments, the invention is not intended to be limited to the specific forms set forth herein. On the contrary, it is intended to cover such alternatives, modifications, and equivalents as can be reasonably included within the scope of the invention as defined by the appended claims.

[0016] In the following detailed description, numerous specific details such as specific orders, structures, elements, and connections have been set forth. It is to be understood however that these and other specific details need not be utilized to practice embodiments of the present invention. In other circumstances, well-known structures, elements, or connections have been omitted, or have not been described in particular detail in order to avoid unnecessarily obscuring this description.

[0017] References within the specification to "one embodiment" or "an embodiment" are intended to indicate that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the present

invention. The appearance of the phrase "in one embodiment" in various places within the specification are not necessarily all referring to the same embodiment, nor are separate or alternative embodiments mutually exclusive of other embodiments. Moreover, various features are described which may be exhibited by some embodiments and not by others. Similarly, various requirements are described which may be requirements for some embodiments but not other embodiments.

[0018] According to one embodiment of the present invention, a method is provided in which a failure of a first link between a network element and an upstream portion of a communications network is detected and a second link (e.g., a group of links which are logically related to the first link) between the network element and a downstream portion of the communications network is responsively disabled to maintain a communications channel between the downstream and upstream portions of the communications network. While the second link is indicated as "down" or "disabled" to one or more network elements of the downstream portion of the communications network, in actuality the link is active and capable of transferring data between the downstream portion of the network and the network element. Disabling the second link however, serves to quickly notify network elements of the downstream portion of the communications network that a failure has occurred on an upstream link.

[0019] Using embodiments of the present invention, a link failure may be propagated within a bounded or predetermined period of time. For example, where embodiments of the present invention are implemented using predominantly special-purpose hardware (e.g., application specific integrated circuits, special-purpose processors, programmable logic devices, or the like) or specific (e.g., vectored) interrupts, a link between a network element and a downstream portion of a communications network may be disabled within a period of time substantially less than or equal to 50 milliseconds of detecting a failure of another link between the network element and an upstream portion of the communications network.

[0020] Using alternative embodiments of the present invention however a greater period of time for link failure propagation may be necessary. For example, where embodiments of the present invention are implemented using predominantly general-purpose hardware (e.g., one or more general-purpose processors) coupled with machine-executable instructions (e.g., data processing system software, firmware, or the like) or generic (e.g., polled) interrupts, a link between a network element and a downstream

portion of a communications network may be disabled within a period of time substantially less than or equal to 2 seconds of detecting a failure of another link between the network element and an upstream portion of the communications network.

[0021] As the link state or link failure information is propagated downstream, downstream network elements with the capability to switch to alternate (e.g., redundant) links may do so, preserving higher-level communication channels between the downstream and upstream portions of the communications networks and providing seamless failover. Within the present description, the term "downstream" is intended to indicate in a direction from a network's core to a network's edge or towards a network's edge, the term "upstream" by contrast is intended to indicate in a direction from a network's edge to a network's core or towards a network's core. Further within the present description, the term "endstation" is intended to indicate a network element (e.g., a file/data server, personal computer, or other data processing system) residing at the edge of a communications network, the term "switch" is intended to indicate a data link layer network element (e.g., an Ethernet switch), the term "link" is intended to indicate a data link layer connection and may include one or more logical sub-links (e.g., an "Ether-Channel" or "Port-Channel), and the term "failure" is intended to indicate the loss of a data link layer link.

[0022] Fig. 2 illustrates a data processing system including a primary switch network element according to an embodiment of the present invention. As previously described with respect to Fig. 1, data processing system 200 of the illustrated embodiment of Fig. 2 includes an upstream communications network portion 102, a primary switch 104a, a secondary switch 104b, and a number of redundantly-linked network elements (e.g., multi-homed endstations 106a, 106b...106n). Upstream communications network portion 102, (e.g., a network core, wide, metropolitan, or local area network, or the like) including any of a number of network elements is coupled to primary switch 104a using a primary upstream link 108a and to secondary switch 104b using a secondary upstream link 108b.

[0023] Multi-homed endstations 106a, 106b...106n are each similarly coupled to primary switch 104a via one of a plurality of primary downstream links 110a-110n (e.g., via a primary network interface card or host bus adapter) and to secondary switch 104b via one of a plurality of secondary downstream links 112a-112n (e.g., via a secondary network interface card or host bus adapter) as shown. While links 108, 110, and 112

have been illustrated as direct connections between network elements (e.g., upstream communications network portion 102, primary and secondary switches 104, and endstations 106) in alternative embodiments, any of links 108, 110, and 112 may be wireless (e.g., using electro-magnetic, optical, infrared and/or acoustic signals or transmission media). For example, one or more of primary downstream links 110 and secondary downstream links are implemented in one embodiment of the present invention using a wireless local area network (e.g., IEEE 802.11x standard) communication protocol.

[0024] It will be appreciated that the use of the terms "upstream" and "downstream" within the present description is relative based upon the particular network element considered. For example, a link between primary switch 104a and endstation 106a is considered a "downstream" link from the perspective of primary switch 104a and an "upstream" link from the perspective of endstation 106a. Similarly, primary switch 104a may be considered an "upstream" network element from the perspective of endstation 106a and a "downstream" network element from the perspective of upstream communications network portion 102.

[0025] Moreover, a direct physical or virtual link or path need not exist between two network elements for an upstream/downstream relationship to exist and no network element need be coupled to a link for it to be considered "upstream" or "downstream". Consequently, even after primary downstream links 110 are disabled, an endstation 106 is still considered a downstream network element with respect to primary switch 104a.

[0026] In the illustrated embodiment of Fig. 2, primary switch 104a includes a configuration interface 202 and a link failure propagation module 204. While configuration interface 202 and link failure propagation module 204 have been illustrated in Fig. 2 as included within primary switch 104a, in an alternative embodiment of the present invention one or more of configuration interface 202 and link failure propagation module 204 maybe stored at or executed from a switch or other network element within data processing system 200 which is directly or indirectly coupled to one or more of endstations 106a-106n. In one embodiment, configuration interface 202 is used to provide a user interface for configuring link failure propagation module 204 and the operation thereof and link failure propagation module 204 is used to propagate link failure and/or link state as further described herein.

[0027] According to one embodiment, configuration interface 202 is configured to receive data from a user specifying various configuration parameters such as whether or not link failure propagation is to be enabled or not, and if enabled, how link failure propagation is to be performed (e.g., automatically or on demand, for one or more individually specified ports, for all ports associated with one or more specified virtual networks (e.g., virtual local area or storage area networks), or for all ports of a switch). In another embodiment of the present invention, configuration parameters additionally include data specifying what action is to be taken when a previously failed link (or a new link) becomes operational. For example, one or more associated endstations may alternatively "fail back" to a previous primary link which becomes operational again following a failure or the endstation(s) may continue to use a "failed over" secondary link or set of links.

[0028] In operation, data is transmitted between one or more of multi-homed endstations 106a-106n and upstream communications network portion 102 using primary downstream links 110a-110n, primary switch 104a, and primary upstream link 108a. Following a failure of primary upstream link 108a (e.g., due to failure of the physical link hardware, a network interface card, or primary switch 104a) link failure propagation module 204 of primary switch 104a is used to detect the failure and responsively disable one or more of primary downstream links 110a-110n.

[0029] According to the illustrated embodiment of Fig. 2, all downstream links or ports of primary switch 104a are disabled in response to the detection of a failure of primary upstream link 108a. In another embodiment, only those downstream links or ports which are individually predetermined or identified using configuration interface 202 are disabled, and in yet another embodiment, only those downstream links or ports associated with a virtual network such as a VLAN or VSAN which was predetermined or identified using configuration interface 202 are disabled. To disable one or more links according to embodiments of the present invention, any of a number of techniques may be implemented. For example, according to one embodiment of the present invention, a physical layer protocol circuit or "PHY" associated with one or more ports and/or links is disabled using an administrative command, causing all ports and/or links associated with the PHY to become disabled.

[0030] While the illustrated embodiment of Fig. 2 includes a single primary upstream link 108a and secondary upstream link 108b, in alternative embodiments of the present

invention, multiple links between upstream communications network portion 102 and primary switch 104a and/or secondary switch 104b may be used. According to one embodiment, two or more primary upstream links between primary switch 104a and upstream communications network portion 102 are provided. Such primary upstream links may operate simultaneously or in a fail-over configuration with one another. Consequently, a failure of one or more of the primary upstream links need not cause any of downstream links 110 to be immediately disabled. For example, a downstream link 110 may be disabled only upon the failure of all primary upstream links, upon the failure of a predetermined or defined number or proportion of primary upstream links, or upon available bandwidth between primary switch 104a and upstream communications network portion 102 falling below a predetermined or defined threshold level due to the failure or one or more primary upstream links.

[0031] Once a port and/or link (e.g., one or more of downstream links 110) is disabled, an associated multi-homed endstation may failover to a corresponding secondary downstream link 112 by activating a network interface associated with the secondary downstream link and deactivating a network interface associated with the failed primary downstream link. Data is then transmitted between the multi-homed endstation which has failed-over and upstream communications network portion 102 via an associated secondary downstream link 112, a secondary switch 104b, and a secondary upstream link 108b.

[0032] Using one embodiment of the present invention, a network element (e.g., a datalink layer-capable Ethernet switch) is configured to track the state of a virtual network (e.g., a VLAN or VSAN) on all identified upstream interfaces (e.g., links, ports, interface cards, or the like). In the described embodiment, a virtual network can be available on multiple (but not necessarily all) upstream interfaces and can be associated with one or more downstream link(s). Consequently, downstream link(s) can be disabled when all upstream interfaces associated with a virtual network are disabled or failed (i.e., when there are no more upstream links that are members of a downstream port's virtual network). Any disabled downstream links are then re-enabled when an associated virtual network becomes available on any upstream interface.

[0033] Fig. 3 illustrates a link failure propagation process according to an embodiment of the present invention. While a particular order and number of process flowchart elements has been illustrated in the embodiment of Fig. 3, it should be appreciated that a

greater or lesser number of process elements may be used and that the illustrated order may not necessarily be required. For example, in alternative embodiments of the present invention one or more process operations may be performed simultaneously or in parallel. In the illustrated embodiment, link failure propagation configuration data is initially received (process block 302) (e.g., using a configuration interface such as configuration interface 202 of Fig. 2). Thereafter a determination is made as to whether or not link failure propagation is enabled (process block 304). According to one embodiment, the determination of whether or not link failure propagation is enabled is made based upon received link failure propagation data. If a determination is made that link failure propagation is not enabled, the illustrated process is terminated as shown.

[0034] If a determination is made that link failure propagation is enabled on a system associated with the illustrated process embodiment, a loop is entered in which a failure of any upstream link of the associated network element is detected (process block 306). Once an upstream link failure has been detected, a determination is made whether link failure propagation is to be performed on-demand or automatically (process block 308). Where link failure propagation is to be performed on-demand, a determination is made whether or not the current link failure is to be propagated (process block 310) (e.g., in response to a response to a user prompt or lack thereof, or based upon additional processing or analysis of system or environmental attributes).

[0035] According to one embodiment of the present invention, a determination of whether or not a particular link failure is to be propagated on-demand is made by determining (e.g., using a timer) whether any user intervention has occurred within a predetermined amount of time from the occurrence of the link failure. If no user input or intervention is received within the predetermined amount of time in the described embodiment, a default behavior or action (e.g., propagation of the link failure) is performed. Otherwise, a behavior or action specified by the user input or intervention is performed. Where link failure propagation is to be performed automatically (or on-demand in a particular instance) one or more downstream links of a network element associated with the illustrated process embodiment are disabled (process block 312). Thereafter, or following a determination that a particular link failure instance is not to be propagated on-demand, the aforementioned loop is re-entered in which a failure of any upstream link s detected (process block 306)

[0036] Fig. 4 illustrates a data processing system including a primary Ethernet switch network element according to an embodiment of the present invention. In the illustrated embodiment of Fig. 4, a data processing system 400 is implemented including a core network 402, primary and secondary core routers (404a and 404b), primary and secondary network switches (410a and 410b), and primary and secondary Ethernet switches (412a and 412b) coupled together using a plurality of links and further including a plurality of multi-homed endstations 414a-414n similarly coupled to primary Ethernet switch 412a via one of a plurality of primary downstream links 416a-416n (e.g., via a primary network interface card or host bus adapter) and to secondary Ethernet switch 412b via one of a plurality of secondary downstream links 418a-418n (e.g., via a secondary network interface card or host bus adapter) as shown. Using one or more embodiments of the present invention, a failure such as that illustrated on a link between primary network switch 410a and core router 404a may be propagated downstream to one or more of endstations 414. Endstations 414 may include any of a number of network elements (e.g., management modules, processor blades, processor modules, or the like).

[0037] The foregoing detailed description has set forth various embodiments of the present invention via the use of block diagrams, flowcharts, and examples. It will be understood by those within the art that each block diagram component, operation and/or component illustrated by the use of examples can be implemented, individually and/or collectively, by a wide range of hardware, software, firmware, or any combination thereof.

[0038] The above description is intended to be illustrative of the invention and should not be taken to be limiting. Other embodiments within the scope of the present invention are possible. Those skilled in the art will readily implement the steps necessary to provide the structures and the methods disclosed herein, and will understand that the process parameters and sequence of steps are given by way of example only and can be varied to achieve the desired structure as well as modifications that are within the scope of the invention. Variations and modifications of the embodiments disclosed herein can be made based on the description set forth herein, without departing from the scope of the invention.

[0039] Consequently, the invention is intended to be limited only by the scope of the appended claims, giving full cognizance to equivalents in all respects.